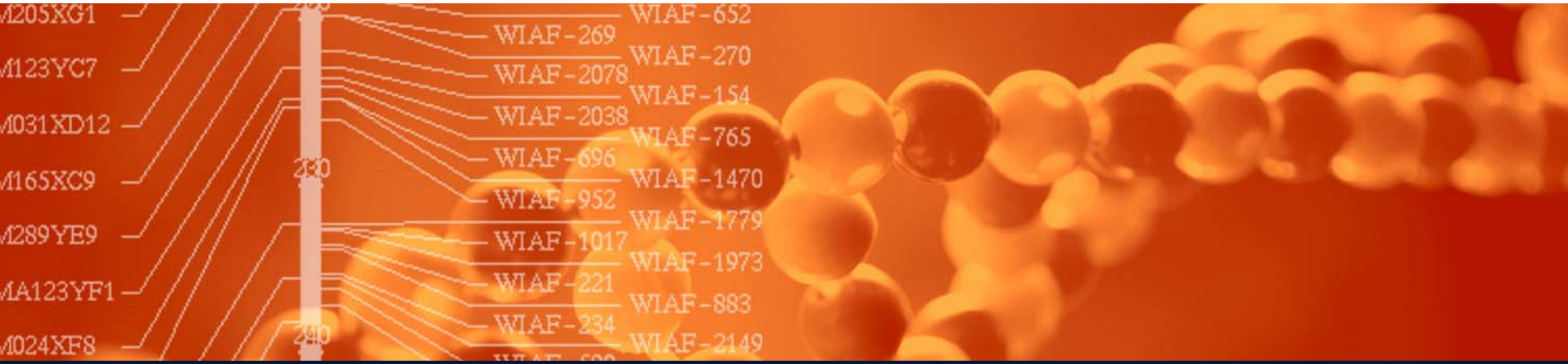
18 February 2004

U.S. GENOMICS



Pathogen Identification By DNA Barcoding

Mark Nadel

© 2003 and confidential US Genomics, Inc.

Outline

- DirectLinear™ Analysis
- DARPA *in silico* Test Scenario
- Technical Framework
- Results

© 2003 and confidential US Genomics, Inc.

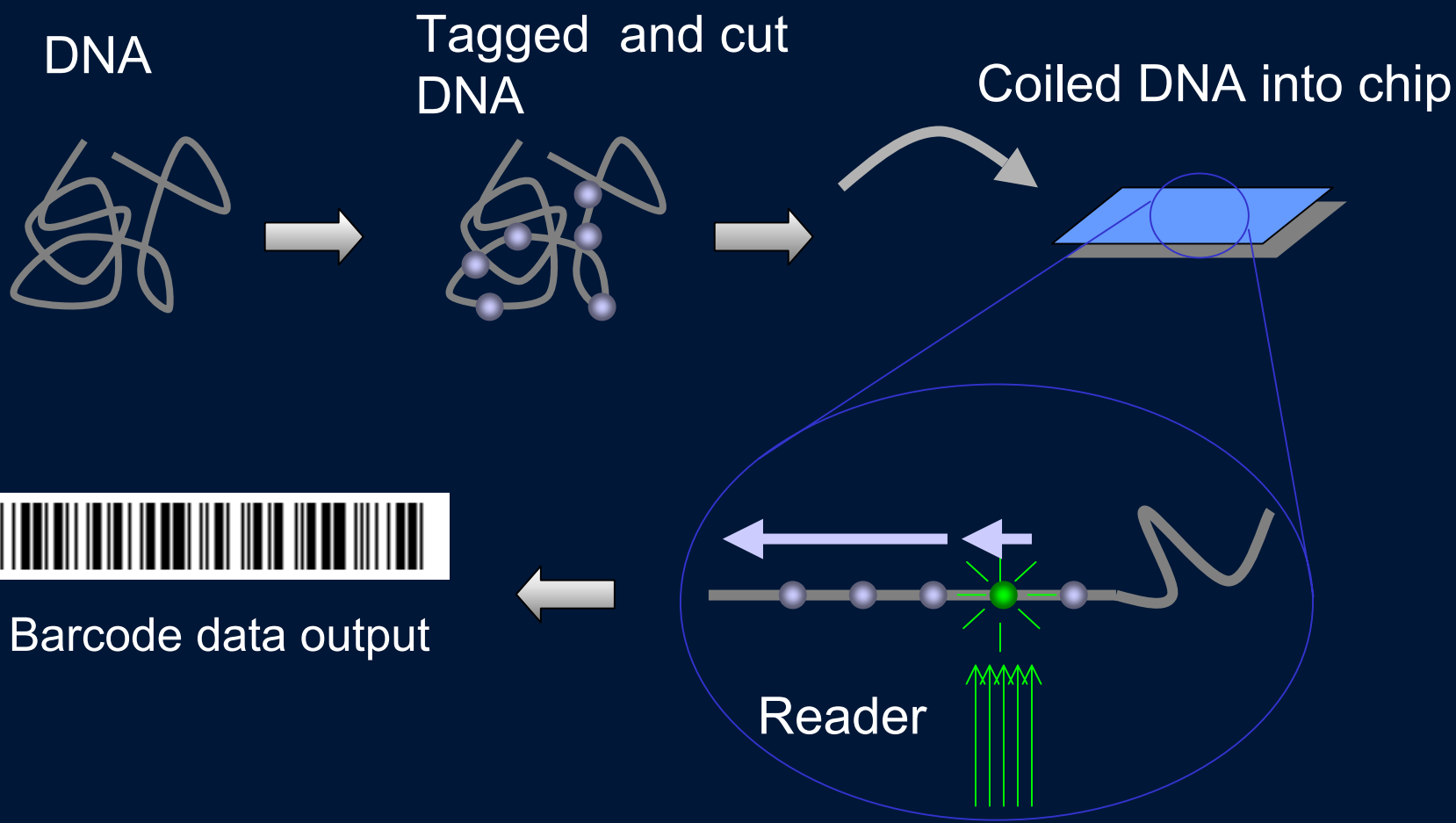
Outline

- **DirectLinear™ Analysis**
- DARPA *in silico* Test Scenario
- Technical Framework
- Results

© 2003 and confidential US Genomics, Inc.

U.S. GENOMICS

DirectLinear™ Analysis

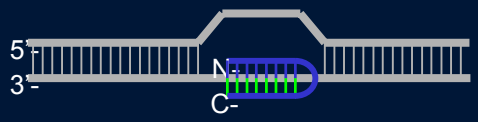


© 2003 and confidential US Genomics, Inc.

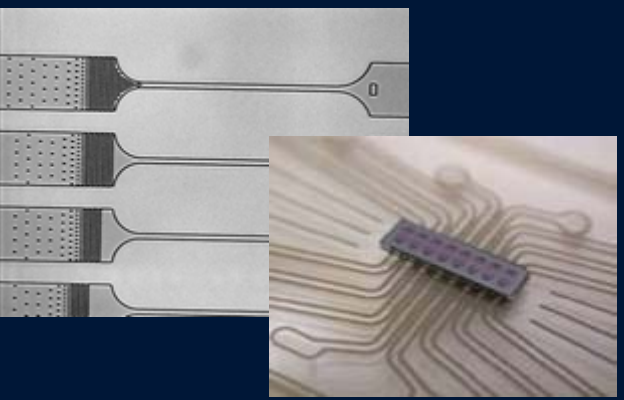
U.S. GENOMICS

GeneEngine™ Platform

Sample prep and tagging

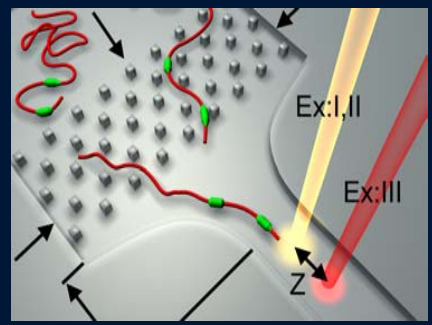


Microfluidics

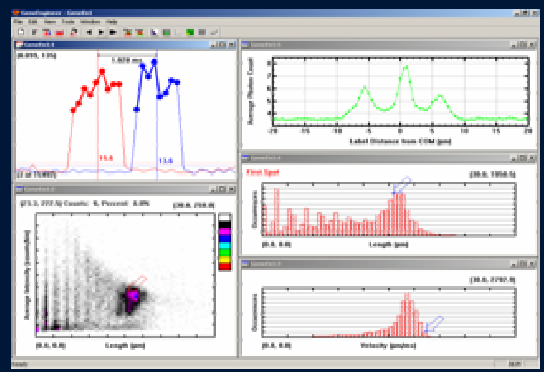


GeneEngine™ Platform

Optics



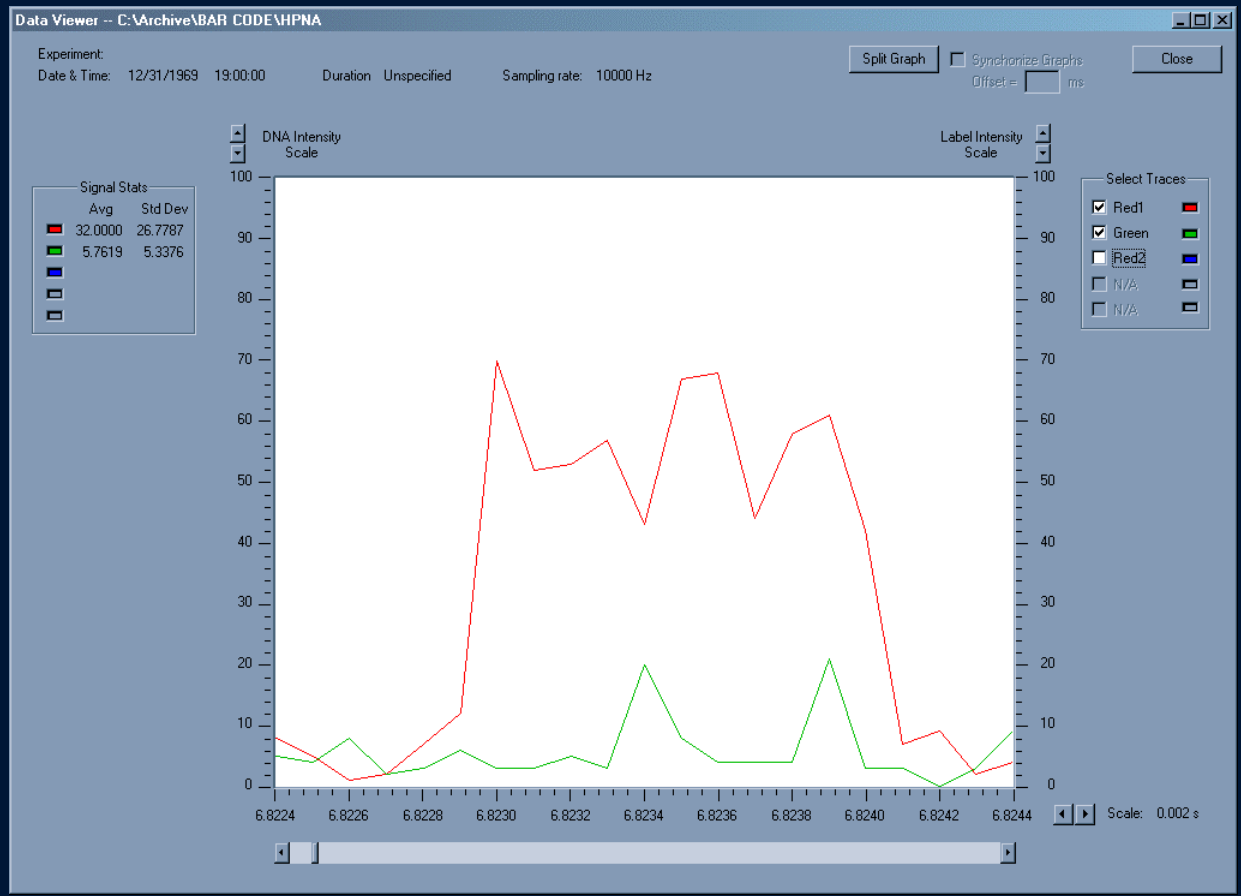
Data capture and analysis



© 2003 and confidential US Genomics, Inc.

U.S. GENOMICS

GeneEngine™ Trace Files



© 2003 and confidential US Genomics, Inc.

For Further Information

Eugene Y. Chan, Nuno M. Goncalves, Rebecca A. Haeusler, Amie J. Hatch, Jonathan W. Larson, Anthony M. Maletta, Gregory R. Yantz, Eugene D. Carstea, Martin Fuchs, Gordon G. Wong, Steven R. Gullans, Rudolf Gilmanshin,

DNA Mapping Using Microfluidic Stretching and Single-Molecule Detection of Fluorescent Site-Specific Tags

Genome Research (2004), in press.

© 2003 and confidential US Genomics, Inc.

Outline

- DirectLinear™ Analysis
- **DARPA *in silico* Test Scenario**
- Technical Framework
- Results

© 2003 and confidential US Genomics, Inc.

In silico Test Scenario

- Targets → Y. pestis, V. cholorae, B. subtilis
- Background → 80 bacteria and yeast species

© 2003 and confidential US Genomics, Inc.

Outline

- DirectLinear™ Analysis
- DARPA *in silico* Test Scenario
- **Technical Framework**
- Results

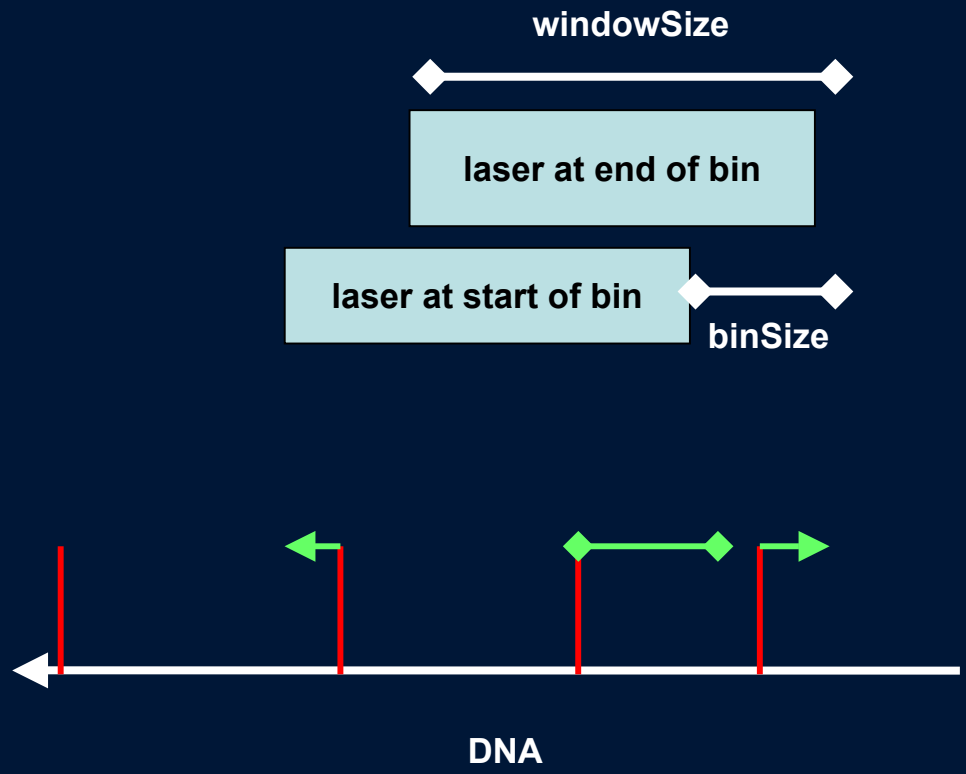
© 2003 and confidential US Genomics, Inc.

Underlying Method

- Fragment DNA to isolate by size
- Tag with a "motif"
- Each fragment has a set of possible "perfect" barcodes
  - Determined by thresholding trace files
  - Motif Map + "parameter settings" + starting point → perfect barcode

© 2003 and confidential US Genomics, Inc.

GeneEngine™ Simulator



© 2003 and confidential US Genomics, Inc.

Basic Test Strategy

- Check for perfect barcodes of certain fragments of targets
- Sound the alarm if we see some predetermined number
- The challenge is to find an enzyme and tag so that each target will have some “identifier” fragment(s) such that:
  - If the organism is present there is a high probability of seeing one of the identifier’s perfect barcodes.
  - If the organism is not present, there is a very low probability of seeing one of the identifier’s perfect barcodes

© 2003 and confidential US Genomics, Inc.

Error Model

- Each fragment has a set of (imperfect) barcodes
  - Perfect sites
  - Single end mismatch (semm) sites.
  - Single interior mismatch (simm) sites.
- Non-specific binding.

Example: Tag site **ATGC**

Perfect	**ATGC**
SEMM	**TTGC**
SIMM	**AAGC**
Nonspecific	**CGAA**

© 2003 and confidential US Genomics, Inc.

Adding Probabilities

- Each fragment has a set of (imperfect) barcodes
  - Perfect sites
    - .95 chance of binding at any specific perfect site.
  - Single end mismatch (semm) sites.
    - .04 chance of binding at any specific semm site.
  - Single interior mismatch (simm) sites.
    - .004 chance of binding at any specific simm site.
- Non-specific binding.
  - .000001 chance of binding at an arbitrary site.

© 2003 and confidential US Genomics, Inc.

Key Notions:

- Tag 2 fragments f and g
- resemble(g,f) = probability that if interrogation of g results in a barcode, it will be a perfect barcode of f
  - Best toward 0
- robustness(f) = resemble (f,f); probability that if interrogation of f results in a barcode, it will be a perfect barcode of f
  - Best toward 1

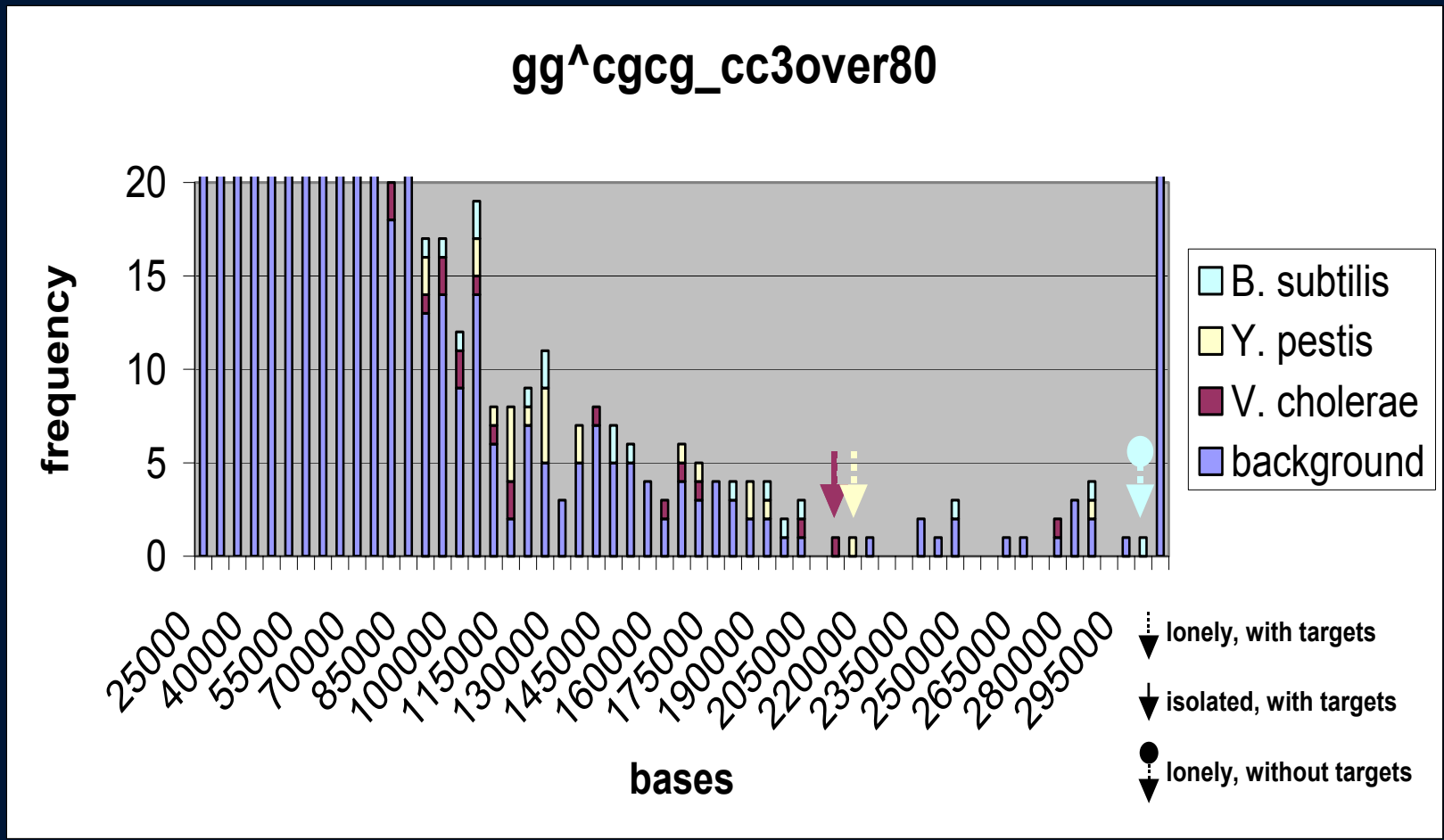
© 2003 and confidential US Genomics, Inc.

Outline

- DirectLinear™ Analysis
- DARPA *in silico* Test Scenario
- Technical Framework
- **Results**

© 2003 and confidential US Genomics, Inc.

Results of Fragmentation by Asc I



© 2003 and confidential US Genomics, Inc.

Identifier Fragment Maps

Fragment at recognition sequence gg^cgcg_cc
Tag at agagagag

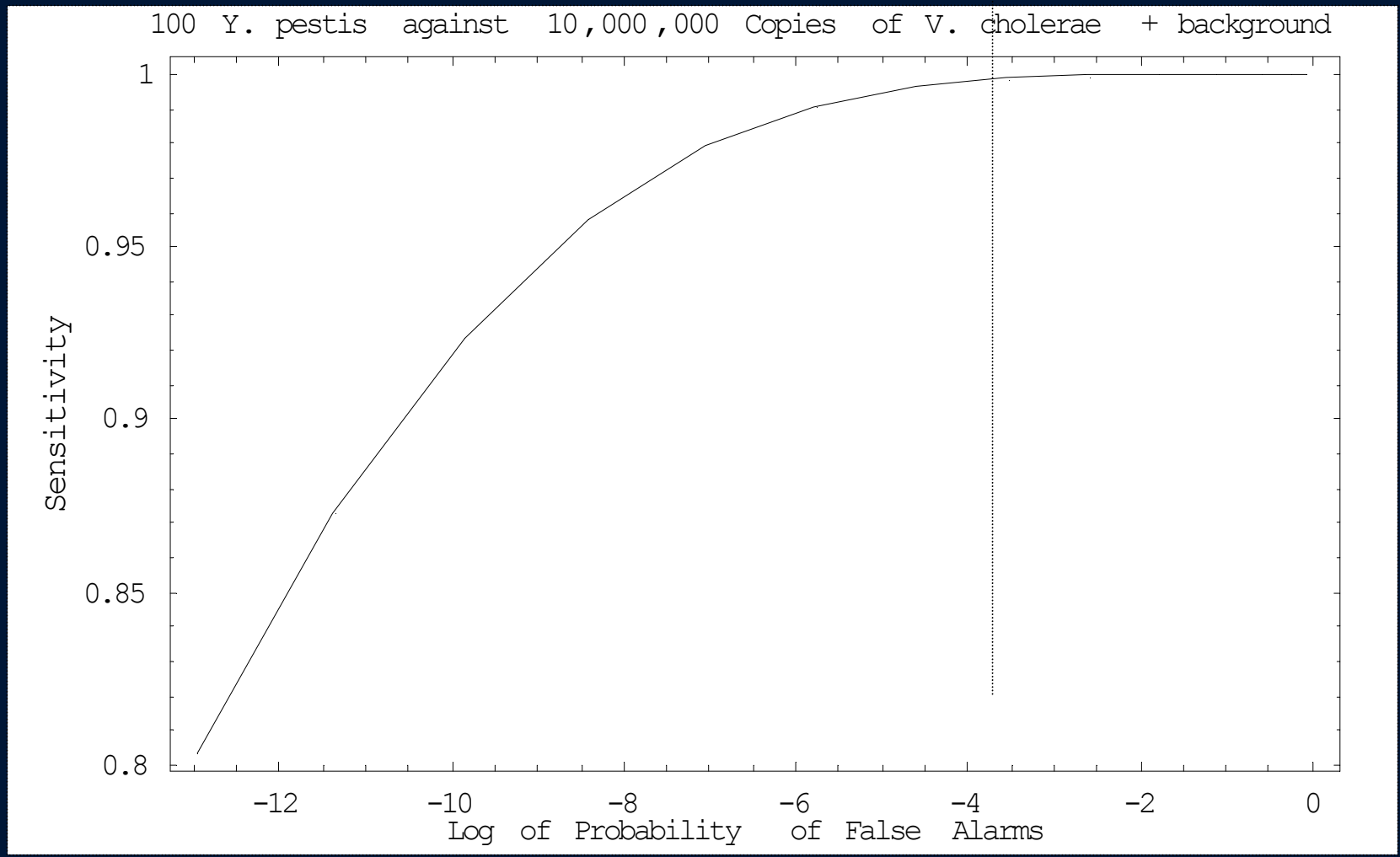
© 2003 and confidential US Genomics, Inc.

Relevant Robustness and Resemble Values

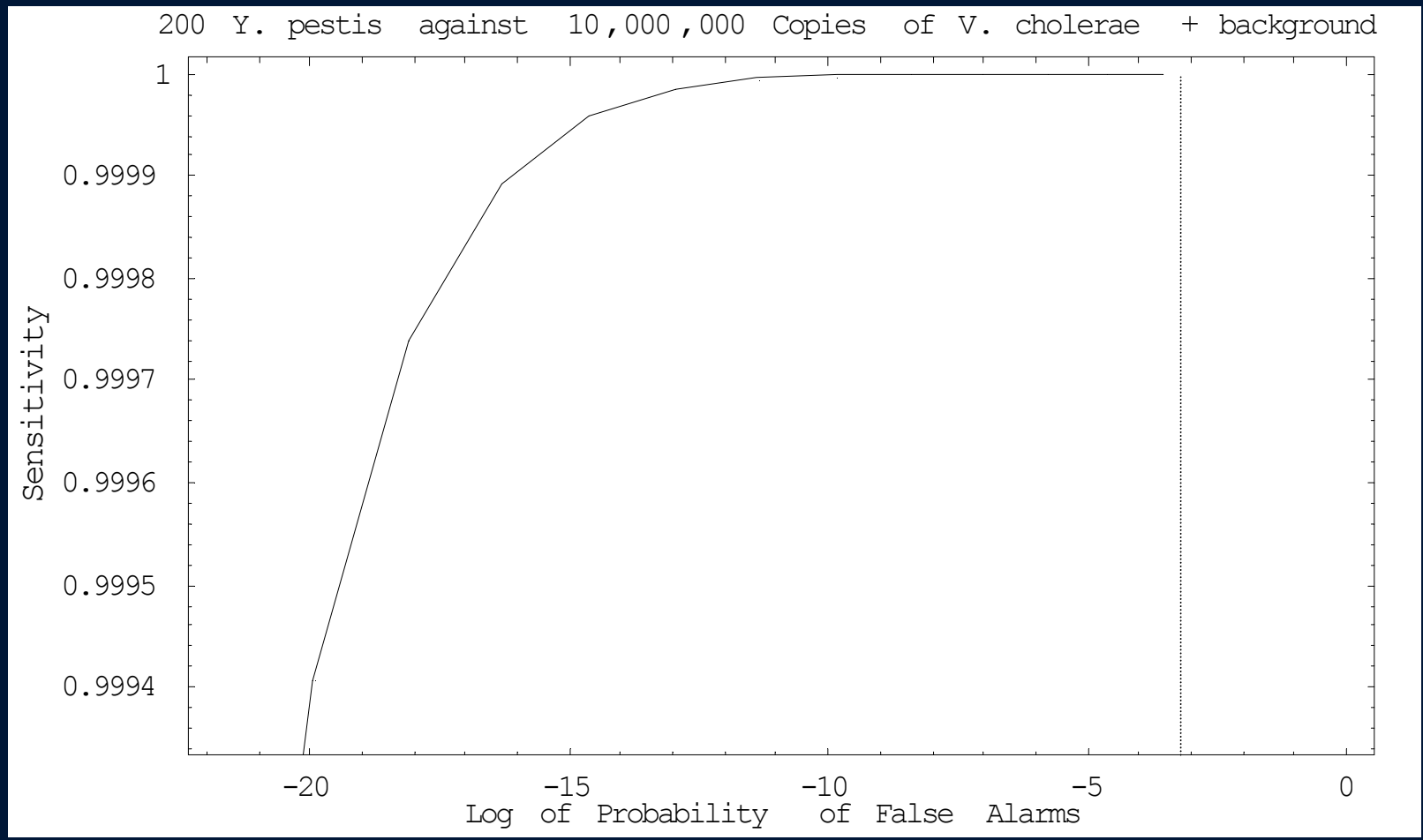
- Robustness of target identifiers
  - robustness(pestis_213,776)=0.557.
  - robustness(subtilis_295,197)=0.319.
  - robustness(cholorae_209,671)=0.307.
- Resemble for potential interlopers
  - resemble(background_218,266, pestis_213,776)= 0
  - resemble(cholorae_209,671, pestis_213,776)= 9.68x10-7
  - resemble(background_291,301,subtilis_295,197)=2.18x10-9
  - resemble(pestis_213,776, cholorae_209,671)=6.42x10-10.
- Use efficiency factor of 0.3
  - Only 30% of fragments from original sample result in a barcode.

© 2003 and confidential US Genomics, Inc.

ROC Curve for Target Y. pestis

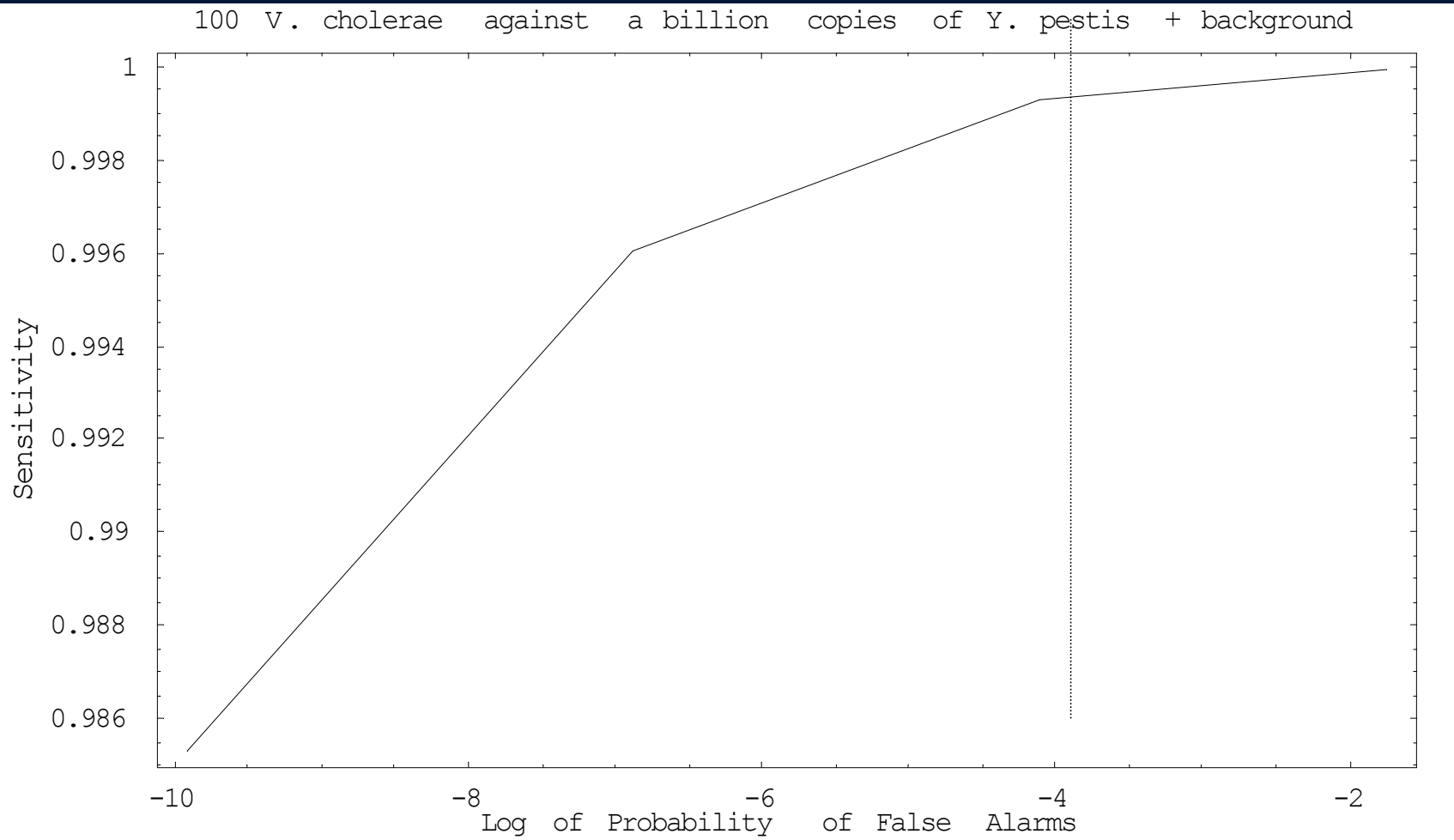


Alternate ROC Curve for Target



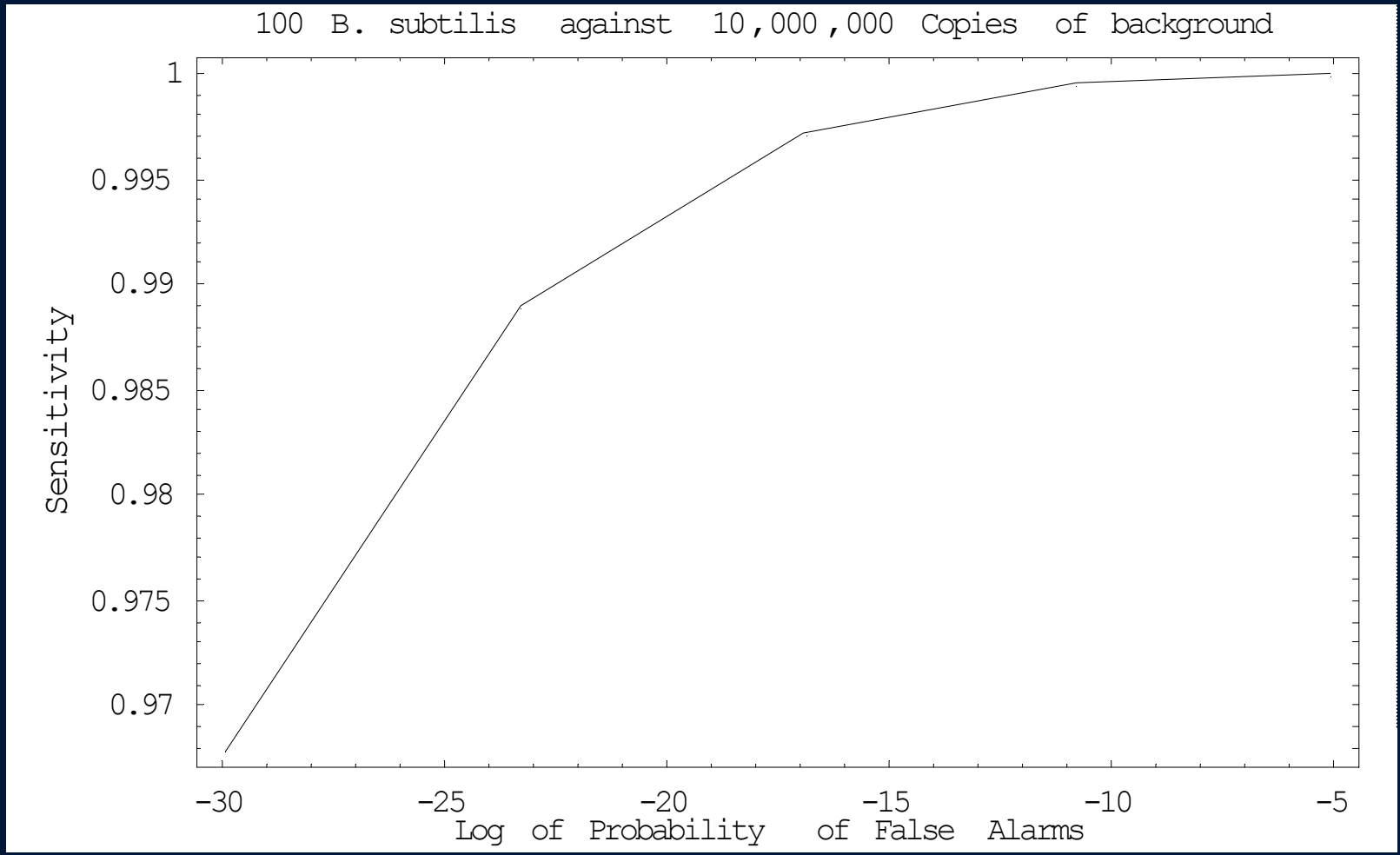
© 2003 and confidential US Genomics, Inc.

ROC Curve for Target V. cholorae



© 2003 and confidential US Genomics, Inc.

ROC Curve for Target B. subtilis



© 2003 and confidential US Genomics, Inc.

Acknowledgements

- DARPA
- Dr. Bin Feng
- Dr. Rudolf Gilmanshin
- John Harris
- Dr. Hai-Ping Ko
- Dr. Michael Shia

© 2003 and confidential US Genomics, Inc.